Data Wrangling for Project 2:

Gather:

1. The first dataset, 'twitter enhanced.csv,' was supplied and only required reading using the pd.read csv() method.
2. The second dataset ‘predictions_df’ was from Udacity's servers, which I had downloaded programmatically using the Requests library from the following URL h_ps://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image predic_ons /image-predic_ons.tsv
3. The json.txt file containing the Twitter query was provided for me because I had trouble requesting a Twitter developer account. Using for loops, I read this.txt file line by line into a pandas dataframe.

Assess: I used inbuilt functions such as head(), tail(),sample(), describe(), info(), isna() among others to assess all three datasets.

Clean: I was able to clean 8 quality issues and 2 tidiness issue as listed below:

Quality issues

These were discovered using visual and programmatic assessment.

Visual assessement

1. The twitter archive dataframe contains null value which are represent both 'NaN' and 'None' in doggo, floofer, pupper, and puppo features.
2. Feature names p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog in the image_pred dataframe are not descriptive.
3. Dog names in features p1 and p2 start with a mixture of upper and lower case in the image_pred dataframe.

Programmatic assessement

4. The tweet_id data type is inconsistent in all dataframes. It is integer in the twitter archives and the image predictions dataframes while its of type object in the tweet_json dataframe which calls for standardization.
5. In the tweet_archive dataframe, the 'timestamp' and 'retweeted_status_timestamp' feature data type is object which needs to be changed to timestamp in order to use these two features in our analysis.
6. Features like "in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls" have missing values and such they need to be cleaned. All these features except expanded_urls have more than 90% of the total observations missing. Perhaps this means we discard them.
7. We have accented characters in the tweet_archive dataframe. An example in case is in the text feature. These need to be handled.

8. Looking at the rating_numerator feature in the tweet_archive, a couple of data points appear to be unrealistic with respect to the rating_denominator. With the help of regular expressions, we can see that some numerator values are listed in the text field.

Tidiness issues

1. In tweet_archive dataframe doggo, floofer, pupper, puppo features are dog names. Therefore, after sorting out the quality issue associated with these fields, they can be dropped
2. There is an intersection between all dataframes and as such we shall need to merge them after cleaning.